Paraphrased Plagiarism Detection Using Sentence Similarity

Zubarev D.V. – PhD student
Sochenkov I.V. – PhD

+7 (499) 135-04-63
117312, Moscow
pr. 60-letiya Oktyabrya, 9
Plagiarism detection task

- Two subtasks
  - Source retrieval - given a suspicious document and a large collection of sources, the task is to retrieve all plagiarized sources while minimizing retrieval costs
  - Text alignment - given a pair of documents, the task is to identify all contiguous maximal-length passages of reused text between them
Related work

• Source retrieval
  • Querying search engines
  • Methods revolve around selecting keywords
  • Many heuristics for candidates filtering

• Text alignment
  • Many methods exist based on N-grams, skip N-grams, syntactic N-grams, stop words N-grams
  • Vector space models with cosine similarity are also widely used

• There was competition PAN (2009-2015)
Preprocessing of collection of sources

- Index all sources for future fast retrieval
- Store extra information about each word (PoS-tags, semantic roles, etc.)
- Some statistics of a source collection:
  - 5.7 million texts
  - 130 GB – raw size
  - 229 GB – size of indexes
Document preprocessing: Linguistic analysis

- Perform deep natural language processing of the uploaded text
  - POS-tagging
  - Syntactic parsing
  - Semantic role labeling
  - Semantic relation extraction

Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts, Dialogue 2014
First stage: Candidates retrieval (source retrieval)

- Employ Vector Space Model and modified Hamming distance
- Use some noun phrases along with words for creating a vector
- Words and phrases are weighted (TF-IDF)
- Only top 100-200 are used
- The 600 most similar documents are retrieved on this stage
- They are called candidates
Second stage: Suspicious sentences selection

- Filter sentences based on various criteria:
  - a TF-IDF weight of a sentence
  - a length of a sentence
  - an amount of non-alphanumeric symbols in a sentence

- TF-IDF weighting schema is used
- IDF weights are calculated based on word frequencies in all collections
- Top 10000 weighted sentences are selected
Second stage: Fast filtering° (Text alignment)

- Intersect each selected sentence from the suspicious document with all other sentences from the candidates.
- Apply fast algorithm for estimation of the size of intersection for filtering most irrelevant sentences with unmatched lexis.
- Pairs of sentences that share at least 35% of words are passed to the next stage.
Third stage:
Sentence similarity (Text alignment)

• Calculate multiple similarities of each pair using different measures:
  • lexis similarity measure
  • syntactic similarity measure
  • semantic similarity measure
• Combine each obtained value into overall similarity
• Pairs that exceed predefined similarity threshold are considered to be incorrectly reused fragments
Tuning plagiarism detection method (Random search)

- 13 parameters to tune:
  - each of them has from 10 to 20 values
- Initialize each parameter with random value
- On each iteration
  - Slightly tweak each parameter by increasing/decreasing its value
  - Measure performance
  - Choose the best combination
  - Repeat
- Interrupt when the performance of the detection method is not changed for a while
- Repeat the whole search with a new seed
Evaluation corpus from PlagEvalRus 2017

• **Source retrieval:**
  • Sources collection: 5.7 million documents
  • training set: 671 suspicious documents
  • Test set: 10k suspicious documents

• **Text alignment:**
  • training set: 9k pairs
  • Test set:
    • ~10k pairs
    • available only on evaluation platform Tira
Evaluation corpus (2)

- Evaluation corpus includes plagiarism cases with various obfuscation types:
  - **Essay-1** – manually written essays with plagiarism; copy-paste and light/moderate modifications (only in training dataset)
  - **Essay-2** – manually written essays with plagiarism; moderate/heavy modifications
  - **Generated texts** – texts with randomly generated plagiarism; copy-paste or moderate modifications
  - **Academic texts** – real world examples of plagiarism; mostly copy-paste
Performance Measures (Source retrieval)

- Recall – the fraction of sources that are retrieved
- Precision – the fraction of retrieved documents that are true sources
- Mean average precision (MAP) – the higher the more sources are in the top of the result
Evaluation of source retrieval algorithm

- Results on the test data for source retrieval

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Mean average precision</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academic</strong></td>
<td>0.978</td>
<td>0.61</td>
<td>0.003</td>
</tr>
<tr>
<td><strong>Essays-2</strong></td>
<td>0.989</td>
<td>0.39</td>
<td>0.009</td>
</tr>
</tbody>
</table>
Performance Measures (Text alignment)

• Recall – the fraction of a source text that is detected
• Precision – the fraction of detected text that is plagiarised
• Granularity reflects the consistency of detected text (the less the better)
• Plagdet – the combination of previous three measures
## Evaluation of text alignment

- Results on the test data for source retrieval

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Granularity</th>
<th>Plagdet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essays-2</td>
<td>0.531</td>
<td>0.82</td>
<td>1.0016</td>
<td>0.644</td>
</tr>
<tr>
<td>Baseline: Essays-2</td>
<td>0.076</td>
<td>0.896</td>
<td>1.141</td>
<td>0.128</td>
</tr>
<tr>
<td>Generated paraphrasing</td>
<td>0.865</td>
<td>0.981</td>
<td>1.483</td>
<td>0.7</td>
</tr>
<tr>
<td>Baseline: generated paraphrasing</td>
<td>0.833</td>
<td>0.97</td>
<td>3.464</td>
<td>0.416</td>
</tr>
<tr>
<td>Generated copy/paste</td>
<td>0.859</td>
<td>0.978</td>
<td>1.466</td>
<td>0.702</td>
</tr>
<tr>
<td>Baseline: generated copy/paste</td>
<td>0.994</td>
<td>0.961</td>
<td>1.004</td>
<td>0.9744</td>
</tr>
</tbody>
</table>
## Most difficult obfuscation types for our method

- Training data was annotated with the type of obfuscation
- Recall per type for Essays-1 collection

<table>
<thead>
<tr>
<th>Description</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCT</td>
<td>0.41</td>
</tr>
<tr>
<td>HPR</td>
<td>0.44</td>
</tr>
<tr>
<td>SSP</td>
<td>0.65</td>
</tr>
<tr>
<td>LPR</td>
<td>0.78</td>
</tr>
<tr>
<td>ADD</td>
<td>0.85</td>
</tr>
<tr>
<td>DEL</td>
<td>0.85</td>
</tr>
<tr>
<td>CPY</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Evaluation of the plagiarism detection method

- Results on training data

<table>
<thead>
<tr>
<th>Source Retrieval</th>
<th>Text Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec.</td>
<td>MAP</td>
</tr>
<tr>
<td>Essays-1</td>
<td>0.97</td>
</tr>
<tr>
<td>Essays-2</td>
<td>0.82</td>
</tr>
</tbody>
</table>

- Results on test data

<table>
<thead>
<tr>
<th>Source Retrieval</th>
<th>Text Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec.</td>
<td>MAP</td>
</tr>
<tr>
<td>Essays-2</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Future work

- Estimate current impact of semantic/syntactic similarity measures on recall
- Explore more possibilities to leverage them for detecting heavily disguised plagiarism
- Address weak points of detection some obfuscations (concatenation)
Denis Zubarev - zubarev@isa.ru
Demo - like.exactus.ru